

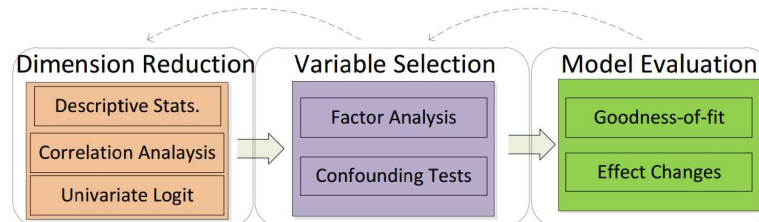
读 A Visual Analytics Approach to High-Dimensional Logistic Regression Modeling and its Application to an Environmental Health Study 一些感想

文章做的是依据 logistic 回归模型的 pipeline, 设计可视分析系统. 其研究的数据是婴儿出生缺陷 (相当于结果变量, y) 和环境 (相当于一系列的解释变量, x_i) 的关系. 在分析中, 主要的几个挑战是: 过拟合, [干扰项](#), [多重共线性](#), 弱效应. 本文重点也就是解决这几个问题.

文中相关工作也提到了我们之前在综述中提到的系统比如 INFUSE 以及 a partition-based framework for regression model building. 但是指出有这些不同之处: 1) 本文解决的重点是如上四个; 2) 度量变量选择的好坏, 模型好坏的指标需要更好的; 3) 本文更偏重于解释性的回归建模, 而上述的都偏重于预测性的建模 [这和数据挖掘的类似]. 第 3) 点是比较重要的, 因为作者表示预测性建模总是在最小化 prediction error, 而解释性的侧重于特征表达, 和找到变量之间的关系以找到真的风险因素 (即是真正与结果变量正相关的解释变量). 整个分析流程中蕴含着这一点.

文中出现了大量专业术语, 尽管之前有公安那边的生存模型和今年投稿的回归模型, 以及以前看过的一些可视化文章, 我没见过这么多专业术语和方法并且从头到尾贯穿. 一方面这可能是我在这一块学的比较浅; 一方面本文可能亮点就在这个非常复杂且专业的统计学基础上, 因为粗看这些可视设计真的没有太大创新性可言, 交互上也是按照 pipeline 流水线下来的, 在每个模块内部能做到联动, 重点高亮, 相关信息提示等等, 很细致, 也不算新颖.

整个可视分析的流程如下图:



简单点说, 第一步是对单一解释变量进行分析, 找出来有用的; 第二步是对两两解释变量之间的关系进行分析; 第三步骤是针对模型建立进行分析. 这个步骤也是可以循环的, 如果不满意可以回退到之前的步骤. 这样一个流程, 和那个 partition-based framework 也是神似的.

每一步中, 对那些检验结果优劣的变量方法都有可视以及交互的编码, 几乎可以说面面俱到, 便于选择; 对数据分布情况, 采用基于像素的编码方式, 个人赞同这是一个很好的选择, 用紧凑的空间详细展示了数据的属性, 排布采用一种螺旋式的结构也比较科学. 我联想我们三月份投稿时候那个像素图, Andrienko 指出了我们那样从高到低依次排会导致刷选时候意义上的不连贯; 这是一个解决方法. 从我们当时选这个像素图本身来说, 也并没有本文用像素图那么有说服力.

整个流程, 如果真的要完成一套任务, 其实仍需要不少的时间精力, 很多地方需要手动不断的调整, 需要一对对比较. Case study 中有一个 86 变量的任务, 然而并没有提及时间效率我也不好下定论. 但是一般这样的任务, 如果纯用自动方法, 常常会结果不理想; 并且一般维度都不会特别大, 相对来说人工花时间在上面如果结果明显优于纯机器自动方法那也是值得的. 这里便又有那个问题, 人在这样的计算机的分析过程中到底要占多大的比重才

合适.

文章写作方面,思路比较清晰,对概念解释,前后连贯和一致性上不算太完备不过也大致 OK (当然这可能还是我水平问题).想起我们投稿时,变量和术语比较多,一开始写着写着就乱了,最后关头还有些瑕疵.但是文中涉及统计专业知识的写着没有乱,该提到都有提到,很细致,这一点值得学习.

文章这种方法能不能借用到我现在已知的一些项目.我个人觉得还不能,其实这篇文章标题好像和健康有关,分析流程中很少有提到相关内容.这一点,其实上面相关工作提到的也一样有这样的情况,他们的方法在这个领域内由于数据类型的相似,可以运用;如果换了一个领域可能又得做另外的深入研究了.之前不管是何英华老师还是彭泰权老师,在做的项目中都没有提到说要把类别性变量单独拿出来的,本文就是这么做的;选择变量的时候我们也没考虑到这么复杂的变量间关系.当然这个可能又和具体做的原始驱动有关,正如上面提到的解释性和预测性回归的差别.如果我们要用过来,要么就是针对我们做的流程做一个定制和展示,要么在别的环节上有所创新.

最后我还想说由于本身水平有限,如果作者在文中提到的那些度量结果优劣的指标和一些方法,如果本身就是不合理的,我也无法识别.文章理论基础很强,对这个方法流程每一步忠实的可视化,物尽其用,是我认为的本文的最大优点.

陆俊华